计算机科学
COMPUTER SCIENCE

基于联邦学习的Gamma回归算法

郭艳卿，李宇航，王湾湾，付海燕，吴铭侃，李祎

引用本文

郭艳卿，李宇航，王湾湾，付海燕，吴铭侃，李祎基于联邦学习的Gamma回归算法[J]．计算机科学，2022，49(12)：66-73.

GUO Yan-qing，LI Yu-hang，WANG Wan-wan，FU Hai-yan，WU Ming-kan，LI Yi. FL-GRM:Gamma Regression Algorithm Based on Federated Learning [J]. Computer Science，2022，49(12)：66-73.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220600034

基于联邦学习的 Gamma 回归算法

郭艳卿[1]　李宇航[1]　王湾湾[2]　付海燕[1]　吴铭侃[1]　李　祎[1]
1 大连理工大学信息与通信工程学院　辽宁 大连 116024
2 深圳市洞见智慧科技有限公司研究中心　北京 100028
(guoyq@dlut.edu.cn)

摘　要　在水文学、气象学以及保险理赔评估等领域中,通常假设因变量服从 Gamma 分布,相比多元线性回归,在 Gamma 分布假设下建立起的 Gamma 回归具有更出色的拟合效果。以往获得 Gamma 回归模型的方法是将数据集中起来进行训练,当数据是由多方提供时,在不交换数据的情况下训练满足隐私保护的 Gamma 回归模型成为需要解决的问题。为此,提出了一种多方安全的纵向联邦 Gamma 回归算法,该算法首先使用迭代法推导出纵向联邦 Gamma 回归模型的对数似然估计表达式,然后结合工程实际确定模型的连接函数,进而构造损失函数建立参数的梯度更新策略,最后对同态加密后的各方参数进行融合更新,获得联邦学习后的 Gamma 回归模型。在两种公开数据集上进行性能测试,实验结果表明,所提联邦 Gamma 回归算法在不交换数据的前提下,可有效利用多方数据的价值生成 Gamma 回归模型,该模型对数据的拟合效果逼近数据在集中情况下学习到的 Gamma 回归模型,优于单方独立学习获得的 Gamma 回归模型。

关键词:联邦学习;Gamma 回归;同态加密;隐私保护;多方安全计算
中图法分类号　TP181;TP309

FL-GRM:Gamma Regression Algorithm Based on Federated Learning

GUO Yan-qing[1],LI Yu-hang[1],WANG Wan-wan[2],FU Hai-yan[1],WU Ming-kan[1] and LI Yi[1]
1 School of Information and Communication Engineering,Dalian University of Technology,Dalian,Liaoning 116024,China
2 Research Center of InsightOne Tech Co.,Ltd.,Beijing 100028,China

Abstract　People commonly hypothesize that an independent variable follows a Gamma distribution in many areas,including hydrology,meteorology and insurance claim. Under the Gamma distribution assumption,Gamma regression model enables an outstanding fitting effect,compared with multivariate linear-regression model. Previous studies may be able to obtain a Gamma regression model trained only on a public dataset. However,when the datasets are provided by multiple parties,how to seek to address the problem of data privacy by training Gamma regression model without exchanging the data itself? A secure multi-party federated Gamma regression algorithm has been applied to this area. Firstly,the log-likelihood function is derived with the iterative method. Secondly,the link function is determined according to the fact,and the gradient updating strategy is constructed by the loss function. Finally,the parameters with homomorphic encryption are updated,then the training is completed. The model is tested on two public datasets,and the results show that under the premise of privacy protection our method can effectively use the value of multi-party data to generate Gamma regression model. The fitting performance of our method is better than that of Gamma regression model implements in a single part,and is close to the result yielded by centralized data learning model.

Keywords　Federated learning,Gamma regression,Homomorphic encryption,Privacy protection,Secure multi-party computation

1　引言

回归模型是一类重要的统计工具,它使用连接函数在自变量和因变量之间建立起联系。但是在实际应用中,相应变量并不都满足其假定的条件,因此 Nelder 等于 1972 年提出了广义线性模型[1]。当因变量服从 Gamma 分布时,广义线性模型即为 Gamma 回归模型(Gamma Regression Model,GRM)。当数据来自多个参与方时,传统机器学习将多方

到稿日期:2022-06-06　返修日期:2022-08-29
基金项目:国家自然科学基金(62076052,62106037,U1936117);中央高校基本科研业务费(DUT20TD110,DUT20RC(3)088);国家社科基金重大项目(19ZDA127);模式识别国家重点实验室开放课题项目(202100032)
This work was supported by the National Natural Science Foundation of China(62076052,62106037,U1936117),Fundamental Research Funds for the Central Universities(DUT20TD110,DUT20RC(3)088),Major Program of the National Social Science Foundation of China(19ZDA127) and Open Project Program of the National Laboratory of Pattern Recognition(NLPR)(202100032).
通信作者:付海燕(fuhy@dlut.edu.cn)

数据传输到云端，在集中的数据集上训练获得 Gamma 回归模型，但这种数据传输方式会产生隐私泄露问题。近年来，世界各国越来越重视对数据隐私的保护，针对隐私保护的法律也陆续出台，来自不同机构或个人的原始数据不能被任意收集和使用。这些法律法规的约束导致了数据孤岛的产生，数据源之间不能够进行数据交互，这使得通过数据集中进行回归模型训练的传统学习方式变得不可行。

为了解决数据孤岛问题，联邦学习提出“数据不动模型动”的思想，将多方的模型进行融合优化。各参与方无须传递和共享原始数据，在数据不出本地的情况下，即可进行数据的联合训练和应用，建立合法合规的机器学习模型。

为了解决多参与方共同学习 Gamma 回归模型所产生的隐私泄露问题，本文研究了纵向联邦学习中 Gamma 回归模型的建立以及参数更新方法，提出了一种多方安全的纵向联邦 Gamma 回归算法。该算法首先使用迭代法推导出纵向联邦 Gamma 回归模型的对数似然估计表达式，然后结合工程实际确定模型的连接函数，进而构造损失函数建立参数的梯度更新策略，最后对同态加密后的各方参数进行融合更新，获得联邦学习后的 Gamma 回归模型。在模型交互过程中，由于采用同态加密技术对不同参与方之间传递的数据进行加密，因此能够保证各方数据的隐私安全。

本文的主要贡献如下：

(1)首次提出了针对 Gamma 回归模型的纵向联邦学习框架；

(2)采用对数连接函数增大模型适用范围并优化损失函数；

(3)提出了一种基于同态加密的多方协同参数更新算法。

2 相关工作

2.1 Gamma 回归

1972 年广义线性模型被提出后，研究者们在该模型下进行了很多拓展工作[1-2]。在广义线性模型中，当因变量服从 Gamma 分布时，其被称为 Gamma 回归模型[3]。随后，Gamma 回归模型被广泛应用在水文学、气象学以及保险理赔评估中。

在水文学和生态学领域中，研究者经常使用 Gamma 分布作为因变量的分布函数。Wu 尝试使用多种分布在不同季节和气候区进行实验，发现 Gamma 分布在冬季干旱区的模拟中具有较大的优势[4]。Ma 等采用 Gamma 分布及空间插值法研究降雨变化情况，分析了 2012 年降雨量与季节变化的趋势以及气温对降雨量的影响[5]。Paynter 等从时间角度使用 Gamma 分布拟合欧洲数十个观测地的单季节的日降雨量，抑或是从空间角度在一定区域内拟合该地区总降雨量的分布情况[6]。

在非寿类保险费率厘定的研究中，由于 Gamma 回归模型具有广义线性模型的可解释性，能让保险公司更容易理解模型的作用原理，也让保险公司更愿意接纳该建模方法。Gong 应用广义线性模型尝试对承包和理赔数据进行建模，分析得到了定价因子的相对风险水平和风险单位的保费，以帮助保险公司进行农业保险定价的工作[7]。Zhong 等对非寿保险样本数据进行 Gamma 回归建模，探究加法模型与乘法模型的参数估计和拟合优度[8]。

2.2 联邦学习

针对在隐私保护法案约束下各行业存在的数据无法共享的问题，谷歌于 2016 年提出联邦学习技术，旨在解决多方联合训练模型的问题[9-11]。联邦学习和分布式计算类似，其本质是一种分布式的机器学习技术，其框架结构如图 1 所示。联邦学习框架由中心服务器节点和多个客户端节点组成，各客户端可以是任意有计算能力的设备(如手机、电脑、物联网设备)。中心服务器节点负责生成加密秘钥和聚合生成全局模型，各客户端在中心服务器节点的协助下通过多次迭代得到最优的本地模型。在以上过程中可以应用多种加密技术来保证各客户端的数据不被泄露。

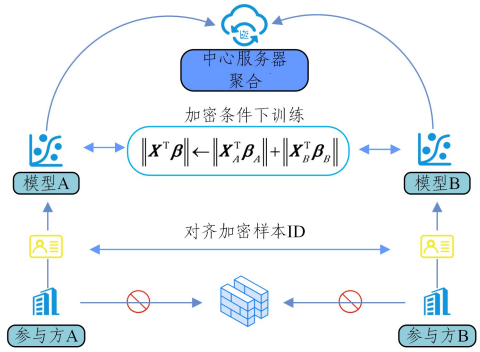

图 1 联邦学习示意图

Fig. 1 Federated learning frameworks

根据参与方数据分布的不同特点，联邦学习可分为横向联邦学习、纵向联邦学习和联邦迁移学习[12-14]。横向联邦学习中各参与方的属性特征重叠较多而用户 ID 重叠较少，典型的应用是手机输入法的下一词语预测[15]。纵向联邦学习是各参与方的用户 ID 重叠较多而属性特征重叠较少，典型的应用是保险理赔、信誉评级等金融属性预测。联邦迁移学习中各参与方的用户 ID 和属性特征都重叠较少，典型的应用是跨国跨平台部门合作。例如中国的汽车销售平台与韩国的电商平台之间的数据迁移，让中国设计的汽车能够更适合韩国消费者的喜好。

越来越多的研究者尝试将联邦学习和机器学习、深度学习算法结合起来[16-18]，以往联邦机器学习相关工作一方面通过选择不同的加密技术对算法进行改进，另一方面通过更改中心服务器结构对算法进行改进[19]。对于联邦逻辑回归模型，Yang 等的工作采用中心化的结构框架，使用同态加密技术进行隐私保护，以迭代的方式对模型梯度进行更新[20]。Yang 等将差分隐私技术与联邦学习相结合，使用去中心的模型训练框架，让标签特征持有方来主导模型梯度更新过程[21]。对于联邦树模型，Liu 等采用中心化的结构框架，以树模型分散储存的方式进行模型训练[22]。Cheng 等提出的特征分桶聚合策略能保证模型预测准确率，同时使用同态加密技术来保护数据隐私[23]。Yang 等在联邦学习框架下设计了具有隐私保护的线性回归模型，其使用的加密技术是同态加密[24]。以往联邦学习的相关工作以基于随机梯度的参数更新方式为主。其中 FedSGD 让参与方将每轮训练后的梯度值上传到中心服务器，服务器聚合后回传给各参与方[25]；Fed-

AVG 让参与方在本地训练多轮，再上传服务器，减少了参与方与服务器之间的通信轮次[11]。在 FedAVG 聚合方式的基础上进行改进的工作有 FedSVRG[26] 和 FedNova[27]。FedSVRG 将各参与方的随机方差缩减梯度作为聚合目标，在相同迭代轮数下使用 FedSVRG 聚合算法得到的模型精度更高；FedNova 从控制本地更新轮数以及全局聚合方式的角度，来提高参与方数据异构时的模型精度。联邦学习攻击与防御也越来越受到研究者的关注。Luo 等设计了特征推断攻击方法来探究纵向联邦学习隐私泄露问题[28]；Wan 等提出了一种攻击自适应聚合策略来防御针对联邦学习框架的攻击手段[29]。联邦学习的最新进展为医疗数据隐私保护技术提供了新思路[30]。Wen 等将联邦学习与区块链技术相结合，将医疗机构声誉值和训练模型存储在区块链上，并利于区块链对医疗机构进行奖励，该方法在提升医疗机构间数据共享效率的同时保护了患者隐私数据[31]。Wang 等将联邦学习技术应用于 COVID-19 胸部 CT 图像分割任务中，同时利于区块链网络替代联邦学习中的中心服务器，解决了医疗数据互不共享以及服务器单点故障的问题[32]。

目前对于联邦回归模型的研究大多以线性回归模型为主。然而 Gamma 分布作为指数分布族中的重要一员，尚未有将其与联邦学习相结合的工作。因此本文提出针对 Gamma 回归模型的纵向联邦学习框架，结合实际工程采用对数连接函数以增大模型的适用范围，并且设计了一种应用同态加密技术的多方协同参数更新算法。

3 联邦 Gamma 回归方法

3.1 符号设置

假设有 n 个样本 $\{\boldsymbol{X}_i, y_i\}_{i=1}^n$，每个样本的属性特征 $\boldsymbol{X}_i \in \mathbb{R}^{1\times d}$ 被分配给 m 个参与方 $\{P_1, P_2, \cdots, P_m\}$，每个参与方拥有的数据 $\boldsymbol{X}^k \in \mathbb{R}^{n\times d_k}{}_{k=1}^m$ 互不重复，且仅参与方 P_1 拥有标签特征 $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$。联邦学习的目的是在 m 个参与方 $\{P_1, P_2, \cdots, P_m\}$ 配合下进行模型的联合训练。

由于每个参与方所拥有的属性特征互不相同，而在进行模型训练之前需要通过加密算法进行用户集合对齐，因此每个参与方所拥有的样本索引 ID 是相同的[33]。本文在纵向联邦学习框架下的符号约定及其含义如表 1 所列。

表 1 符号及其含义

Table 1 Symbols and their meanings

符号	代表含义
m	参与方的数量
n	样本总数
$\{P_i\}_{i=1}^m$	各个参与方
$\boldsymbol{X}^k$	参与方 P_k 拥有的数据
X_j^i	参与方 P_i 的第 j 个属性特征
d_i	P_i 的属性特征数量
d	属性特征总量
$\boldsymbol{y}$	标签特征

3.2 广义线性模型

传统的线性回归模型假设因变量 Y 服从正态分布，其方差为常数，且因变量 Y 与自变量 X 成线性关系。而广义线性模型通常假设因变量 Y 服从指数型分布，自变量 X 通过非线性变换影响因变量 Y 的期望值。广义线性模型包括 3 个部分：随机成分、系统成分和连接函数。

随机成分是因变量 Y 的分布函数，因变量 Y 的每个观察值 Y_i 之间相互独立且服从指数分布族中的一个分布，其概率密度函数表示为：

$$f(Y_i, \theta_i, \phi) = \exp\left(\frac{\theta_i Y_i - b(\theta_i)}{\phi} + c(Y_i, \phi)\right) \tag{1}$$

其中，参数 θ_i 被称为自然参数，ϕ 被称为离散参数，$b(\cdot)$ 和 $c(\cdot,\cdot)$ 根据指数族函数而定。

系统成分 η_i 是自变量 $\boldsymbol{X}_i$ 与模型参数 $\boldsymbol{\beta}$ 的线性组合，可以表示为 $\eta_i = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} = x_1\beta_1 + \cdots + x_d\beta_d$。

连接函数 $g(\cdot)$ 具有单调性和可导性，用来表示随机成分和系统成分之间的关系 $g(E[Y_i]) = g(\mu_i) = \eta_i$，由此可见广义线性模型中，因变量的预测值并没有直接等于自变量的线性组合，而是在自变量的线性组合的基础上进行了一个函数变换。综上所述，广义线性模型的一般表达式为：

$$g(\mu_i) = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} = \sum_j X_{ij}\beta_j \tag{2}$$

3.3 Gamma 回归与参数估计

广义线性模型中连接函数能够代表某种回归模型，如果将 Gamma 回归模型的概率密度函数与广义线性模型中的概率密度函数进行转换，就可以得到 Gamma 回归模型的连接函数。

Gamma 函数 $\Gamma(\alpha)$ 是由阶乘函数扩展得到的，表示为 $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x$。由 Gamma 函数可以得到概率密度函数 $f(y_i, \alpha, \lambda) = \frac{1}{\Gamma(\alpha)\lambda^{\alpha}} y_i^{\alpha-1}\mathrm{e}^{-y_i/\lambda}$，同时令 $\alpha = 1/\phi$，$\lambda = \phi\mu_i$，就可以将 Gamma 回归模型的概率密度函数变换为广义线性模型中概率密度函数的标准形式[34]：

$$f(y_i, \mu_i, \phi) = \exp\left(\frac{-y_i/\mu_i - \ln\mu_i}{\phi} + \frac{1-\phi}{\phi}\ln y_i - \frac{\ln\phi}{\phi} - \ln\Gamma\left(\frac{1}{\phi}\right)\right) \tag{3}$$

其中，均值为 μ_i。利用极大似然估计方法，对式(3)进行连乘运算然后取对数，并结合式(2)，得到 Gamma 分布的对数似然函数式为：

$$L(\mu_i) = \sum_{i=1}^{n}\left(\frac{-y_i/\mu_i - \ln\mu_i}{\phi} + \frac{1-\phi}{\phi}\ln y_i - \frac{\ln\phi}{\phi} - \ln\Gamma\left(\frac{1}{\phi}\right)\right) \tag{4}$$

由于广义线性模型中参数 ϕ 的取值不影响参数估计的结果，为方便计算，令 $\phi = 1$，因此式(4)可以化简为：

$$L(\mu_i) = \sum_{i=1}^{n}(-y_i/\mu_i - \ln\mu_i) \tag{5}$$

3.4 梯度下降法更新模型

为了得到 Gamma 回归模型的连接函数，将 Gamma 回归模型的概率密度函数与广义线性模型中的概率密度函数进行转换，得到 Gamma 回归模型的连接函数关系式 $g(\mu) = -1/\mu$，结合式(2)得：

$$\mu = -\frac{1}{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}} \tag{6}$$

由于式(5)中存在对数函数 $\ln(\mu)$，由对数函数性质可知，μ 的取值范围为 $\mu > 0$，即要求式(6)中 $-\frac{1}{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}}$ 的取值范围为(0,

∞)。然而实际应用中不能保证所有的样本数据的组合值 $-\frac{1}{\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}}$ 都大于 0。因此，我们将 $g(\mu)=\log(\mu)$ 作为新的连接函数代入式(2)，得：

$$\mu_i=\exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}) \tag{7}$$

将式(7)代入到式(5)中得：

$$L(\boldsymbol{\beta})=-\sum_{i=1}^{n}\left(\frac{y_i}{\exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})}+\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right) \tag{8}$$

由于该函数是凹函数，通过取反将其变为凸函数的形式并加入 $L1$ 正则项，得到模型损失函数：

$$Loss(\boldsymbol{\beta})=\sum_{i=1}^{n}\left(\frac{y_i}{\exp(\boldsymbol{X}_i^{\mathrm{T}}\beta)}+\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)+C\sum_{j=1}^{d}\beta_j \tag{9}$$

其中，C 为惩罚因子。对 β_j 求偏导得梯度更新式：

$$\frac{\partial Loss}{\partial\beta_j}=\sum_{i=1}^{n}\left(1-\frac{y_i}{\exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})}\right)X_{ij}+C\,\mathrm{sign}(\beta_j) \tag{10}$$

3.5　参数更新机制

在纵向联邦训练过程中，假定主动方为 P_1，拥有样本标签，可信第三方为中心节点 *Center*，其职责是向所有数据拥有方 $\{P_i\}_{i=1}^{m}$ 发送同态加密所需的公钥 *Key*，设置迭代轮次计数器 *Iter* 自增，并开启第 j 轮迭代。

根据式(10)，对于第 j 批训练数据，每个参与方 $\{P_i\}_{i=2}^{m}$ 计算各自的传播参数：

$$fe_j^i=\exp(-\boldsymbol{X}_j^{i\mathrm{T}}\boldsymbol{\beta}_j^i) \tag{11}$$

其中 fe_j^i 代表每个参与方 $\{P_i\}_{i=2}^{m}$ 向主动方 P_1 传输的数据。使用公钥 *Key* 对传播参数加密后得到 $\|fe_j^i\|$，再将 $\|fe_j^i\|$ 发送给主动方 P_1，即图 2 中步骤①。由于使用了同态加密算法，因此虽然 P_1 得到了其他参与方的加密参数数据，但不能从加密参数中推导出其他参与方的真实数据，故不会造成数据泄露[35]。

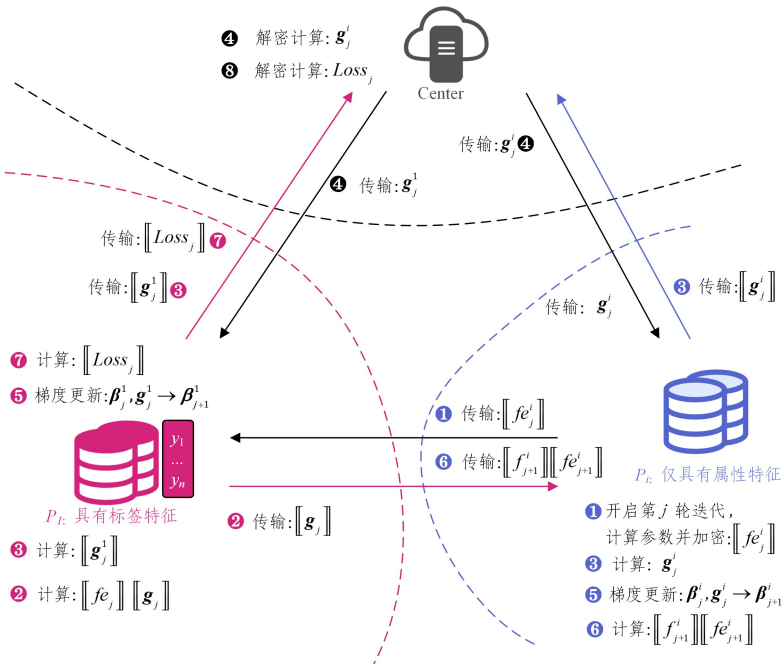

图 2　参数更新过程图

Fig. 2　Diagram of parameter update process

当 P_1 接收完每个参与方 P_i 发送的所有传播参数后，进行模型聚合：

$$\|g_j\|=1-y^1\|fe_j^i\|\exp(-\boldsymbol{X}_j^{1\mathrm{T}}\boldsymbol{\beta}_j^1) \tag{12}$$

然后 P_1 将 $\|g_j\|$ 发送给各个参与方 P_i。为了计算主动方与各参与方的梯度，让 P_1 和各参与方 P_i 使用 $\|g_j\|$ 与各自的第 j 批数据 X_j^i 相乘得到 $\|g_j\|\boldsymbol{X}_j^i$，同时加入 $L1$ 正则项并使用惩罚因子 C 来约束参数大小，即图 2 中步骤②，得到各方的参数 $\boldsymbol{\beta}_j^i$ 梯度更新式：

$$\|\boldsymbol{g}_j^i\|=\|g_j\|\boldsymbol{X}_j^i+C\,\mathrm{sign}(\boldsymbol{\beta}_j^i) \tag{13}$$

数据拥有方 $\{P_i\}_{i=1}^{m}$ 将各自加密的梯度信息 $\|\boldsymbol{g}_j^i\|$ 发送给中心节点 *Center* 进行解密得到 $\boldsymbol{g}_j^i$，然后 *Center* 将 $\boldsymbol{g}_j^i$ 回传给 $\{P_i\}_{i=1}^{m}$，即图 2 中步骤③和④，让它们在学习率 γ 下进行模型参数更新：

$$\boldsymbol{\beta}_{j+1}^i=\boldsymbol{\beta}_j^i-\gamma\boldsymbol{g}_j^i \tag{14}$$

即图 2 中步骤⑤。各参与方根据 $\boldsymbol{\beta}_{j+1}^i$ 计算新的传播参数：

$$\begin{cases} f_{j+1}^i=\boldsymbol{X}_j^{i\mathrm{T}}\boldsymbol{\beta}_{j+1}^i \\ fe_{j+1}^i=\exp(-\boldsymbol{X}_j^{i\mathrm{T}}\boldsymbol{\beta}_{j+1}^i) \end{cases} \tag{15}$$

将 f_{j+1}^i 和 fe_{j+1}^i 加密后发给 P_1，即图 2 中步骤⑥，P_1 计算损失函数：

$$\|Loss_j\|=y^1\|fe_{j+1}^i\|\exp(-\boldsymbol{X}_j^{1\mathrm{T}}\boldsymbol{\beta}_{j+1}^1)+\|f_{j+1}^i\|\boldsymbol{X}_j^{1\mathrm{T}}\boldsymbol{\beta}_{j+1}^1 \tag{16}$$

P_1 将 $\|Loss_j\|$ 发送给 *Center*，*Center* 解密得到 $Loss_j$，并根据 $Loss_j$ 判断当前轮次是否达到模型收敛条件。若没有达到模型收敛条件，则继续进行下一轮迭代；若达到了模型收敛条件，则终止算法迭代，即图 2 中步骤⑦和⑧。算法伪代码如算法 1 所示。

算法 1　FL-GRM 算法

输入：各参与方的数据 $X^1,\cdots,X^m$；标签数据 y；批数量 n；惩罚因子 C；学习率 γ；最大迭代轮次 M；任意小正整数 tol

输出：FL-GRM{联邦 Gamma 回归模型}

1. WHILE Iter<=M DO

2. 中心服务器初始化公钥 Key
3. FOR j=1,…,n DO
4. FOR i=2,…,m in parallel DO
5. $fe_j^i \leftarrow \exp(-\mathbf{X}_j^{i^T}\boldsymbol{\beta}_j^i)$ \\参与方计算传播参数
6. $\| fe_j^i \| \leftarrow \text{Key}(fe_j^i)$ \\加密传播参数
7. 第 i 个参与方向主动方发送 $\| fe_j^i \|$
8. $\| g_j \| \leftarrow 1-y^1 \| fe_j^i \| \exp(-\mathbf{X}_j^{1^T}\boldsymbol{\beta}_j^1)$ \\聚合模型
9. 主动方向第 i 个参与方发送 $\| g_j \|$
10. END FOR
11. FOR i=1,…,m in parallel DO
12. $\| \mathbf{g}_j^i \| = \| g_j \| \mathbf{X}_j^i + \text{Csign}(\boldsymbol{\beta}_j^i)$ \\加密梯度信息
13. 第 i 个参与方向中心服务器发送 $\| \mathbf{g}_j^i \|$
14. 中心服务器解密 $\| \mathbf{g}_j^i \|$
15. 中心服务器向第 i 个参与方发送 $\mathbf{g}_j^i$
16. $\boldsymbol{\beta}_{j+1}^i = \boldsymbol{\beta}_j^i - \gamma \mathbf{g}_j^i$ \\更新梯度
17. END FOR
18. FOR i=2,…,m in parallel DO
19. $f_{j+1}^i \leftarrow \mathbf{X}_j^{i^T}\boldsymbol{\beta}_{j+1}^i$ \\用于计算损失
20. $fe_{j+1}^i \leftarrow \exp(-\mathbf{X}_j^{i^T}\boldsymbol{\beta}_{j+1}^i)$ \\用于计算损失
21. $\| f_{j+1}^i \| \leftarrow \text{Key}(f_{j+1}^i)$
22. $\| fe_{j+1}^i \| \leftarrow \text{Key}(fe_{j+1}^i)$
23. 第 i 个参与方向主动方发送 $\| f_{j+1}^i \|$ 和 $\| fe_{j+1}^i \|$
24. $\| \text{Loss}_j \| \leftarrow \| f_{j+1}^i \| \mathbf{X}_j^{1^T}\boldsymbol{\beta}_{j+1}^1 + y^1 \| f e_{j+1}^i \| \exp(-\mathbf{X}_j^{1^T}\boldsymbol{\beta}_{j+1}^1) + C\sum_{j=1}^{d}\boldsymbol{\beta}_j^i$ \\计算损失
25. END FOR
26. 主动方向中心服务器发送 $\| \text{Loss}_j \|$
27. 中心服务器解密 $\| \text{Loss}_j \|$
28. END FOR
29. Center executers:
30. $\text{Loss} \leftarrow \sum_{j=1}^{n}\text{Loss}_j$ \\计算 n 批数据损失和
31. IF Loss<=tol \\ tol 为任意小正数
32. BREAK
33. ENDIF
34. END WHILE
35. RETURN $\boldsymbol{\beta}$

4 实验

4.1 数据集及评价指标

本文采用 2 个金融保险领域的数据集对联邦 Gamma 回归模型进行性能评估。

数据集 freMTPL2freq 是法国汽车第三方责任索赔数据集，包含 677991 份第三方责任保险单样本，每个样本由 10 维属性特征和 1 个标签组成。

本文调研了麦考瑞大学应用金融与精算学系发布的 12 个保险数据集发现，其中 8 个数据集样本个数均不足 500，还有 3 个数据特征数量不足 5。考虑到纵向联邦学习框架中每一方均要持有不同的特征，特征数量较少的数据集经过分割以后无法使模型学习到必要的知识，因此，该大学发布的数据集中只有数据集 CarData 满足本文的实验要求。

数据集 CarData 来自某年的车辆保险政策，共有 67856 份保险样本，其中 4624 份样本至少有一份索赔，每个样本由 7 维属性特征和 1 个标签组成。

为了验证本文提出的 FL-GRM 方法的有效性，将其与 4 种方法进行实验对比。

LocalA-GRM 和 LocalB-GRM 两种方法的实验设置是仅利用参与方 A 和 B 各自的本地数据进行模型训练，其目的是用来测试非联邦情况下 Gamma 回归模型的效果，验证联邦学习框架的有效性。NoFL-GRM 的实验设置是将全部数据属性特征集中后进行模型训练，即传统情况下的 Gamma 回归方法。其目的是通过与其进行对比，来衡量联邦学习框架下训练的模型的精度损失情况。

FL-LR[25] 是纵向联邦线性回归模型，FL-PRM 是联邦泊松回归模型，通过与它们进行对比来测试 FL-GRM 的数据拟合效果。

本文从模型拟合效果和模型有效性两个角度进行评估。在模型拟合效果方面，首先用 3 个评价指标 Deviance，Log-loss 和 Akaike Information Criterion(AIC)进行模型评价。Deviance 描述的是预测均值与真实值之间的差距，其值越大说明越偏离真实数据。Log-loss 是对数似然比值的负数与样本数量之比，能够排除样本数量对评估结果的影响，其值越小说明模型的拟合效果越好。AIC 衡量模型预测值相比真实值丢失信息的相对量，其值越小说明丢失的信息越少，模型质量越高。有序洛伦兹曲线(Ordered Lorenz Curve)能够对预测结果排序并分箱，通过比较每个分箱内预测均值与真实均值之间的差距，可视化展示模型拟合效果。在模型有效性方面，使用模型损失变化曲线来评估模型的有效性。

4.2 实验结果

将 freMTPL2freq 数据集中每个样本的 10 个属性特征分别按照 2:8，3:7，4:6，5:5 的比例划分给参与方 A 和参与方 B，并将标签特征 y 分配给 A，将其作为主动方，参与方 B 作为协作方。模型 FL-GRM 在 A 和 B 两方的共同参与下进行纵向联邦学习。

实验均采用 $L1$ 正则化，惩罚因子 $C=0.01$，使用批量梯度下降的批大小为 2000，学习率为 $\gamma=0.15$，泊松回归模型参数 $\beta=0.1$。在不同特征分割比例下的实验结果如表 2 所列。

表 2 freMTPL2freq 数据集上不同特征分割比例的实验结果

Table 2 Results of different feature ratios in freMTPL2freq

特征分割比例	模型	Deviance	Log-loss	AIC
2:8	LocalA-GRM	15883.46	9.39	93766.80
	LocalB-GRM	8649.74	8.66	86523.07
	NoFL-GRM	7438.87	8.54	85299.86
	FL-GRM	**10035.16**	**9.01**	**90021.01**
3:7	LocalA-GRM	14804.32	9.30	92762.12
	LocalB-GRM	9107.41	8.89	88933.41
	NoFL-GRM	7438.87	8.54	85299.86
	FL-GRM	**8897.41**	**8.87**	**8887.52**
4:6	LocalA-GRM	14594.23	9.25	92476.43
	LocalB-GRM	9241.37	8.93	89091.37
	NoFL-GRM	7438.87	8.54	85299.86
	FL-GRM	**8834.92**	**8.84**	**88726.31**
5:5	LocalA-GRM	14056.64	9.21	91864.64
	LocalB-GRM	9413.63	8.98	89325.21
	NoFL-GRM	7438.87	8.54	85299.86
	FL-GRM	**8749.23**	**8.82**	**88690.02**

Deviance，Log-loss，AIC 这 3 个评估指标都是越小越好。从表 2 中可以看出，LocalA-GRM 随着特征数量的增多，Deviance，Log-loss，AIC 值都越来越小，其拟合精度越来越好；LocalB-GRM 随着特征数量的减少，Deviance，Log-loss，AIC 值都越来越大，其拟合精度越来越差。两者都弱于 NoFL-GRM 利用全部特征学习的结果，说明随着特征数量的增加，训练出来的模型效果越来越好，也可以证明在单个参与方情况下模型训练的效果与特征数量成正相关。

从表 2 中 FL-GRM 的实验结果可以看出，参与方特征数量的差距大小会对 FL-GRM 的模型性能产生影响。随着两个参与方的特征数量差距减小，FL-GRM 的拟合性能越来越好，但是我们发现，特征分割比例为2∶8 时，FL-GRM 的拟合能力比 LocalB-GRM 的拟合能力差，这是因为 LocalB-GRM 是单方训练并且具有 80%的特征数量，其能够较容易地从这些特征中找到利于提高模型拟合能力的特征，而 FL-GRM 因特征分割比例极度不平衡而未能进行有效的学习。本实验也说明联邦学习中参与方的特征数量差距不宜过大。

我们还进行了多方联邦以及其他特征比例划分的实验测试。对于 freMTPL2freq 数据集，在有 2 个参与方的情况下，将数据集的 10 个属性特征随机划分成 2 份，特征比例为 5∶5，分别分配给参与方 A 和参与方 B。对于有 3 个参与方的情况，各参与方的特征数量分配按照 3 ∶3 ∶4 的比例随机分配。为了方便陈述，用 FL-GRM 表示有 2 个参与方的情况，FL-GRM-3P 表示有 3 个参与方的情况。多个指标的测试结果如表 3 所列。

表 3　freMTPL2freq 数据集上多参与方的模型测试结果

Table 3　Model test results of multiple participants in freMTPL2freq

模型	Deviance	Log-loss	AIC
LocalA-GRM	14056.64	9.21	91864.64
LocalB-GRM	9413.63	8.98	89325.21
NoFL-GRM	7438.87	8.54	85299.86
FL-LR	45498.27	10.84	162287.25
FL-PRM	42179.5	9.53	268777.86
FL-GRM	**8749.23**	**8.82**	**88690.02**
FL-GRM-3P	**8130.73**	**8.63**	**86167.51**

根据表 3 可以看出，FL-GRM 的拟合效果优于 FL-LR，说明当数据集的标签值分布不符合正态分布时，线性回归模型并不能很好地拟合数据。实验结果显示 FL-GRM 与 NoFL-GRM 的测试结果较为接近，因此可以证明 FL-GRM 的模型拟合效果较好。FL-PRM 在 Deviance 和 Log-loss 这两个指标上的结果虽然比 FL-LR 更好，但是其仍然不如 FL-GRM 的拟合效果好。

从表 3 中还可以发现，FL-GRM-3P 与 FL-GRM 的模型效果均优于 LocalA-GRM 和 LocalB-GRM，说明联邦学习能够从多方数据中学习到数据价值。同时对比 FL-GRM 与 FL-GRM-3P 的测试结果发现，后者的模型效果比前者更好。这证明应用本文算法的模型更新方式，即使是在多方参与的情况下，联邦学习得到的模型结果也不会变差。

在数据集 CarData 上，我们进行了有 2 个参与方的实验测试。数据集特征划分比例为 3∶4，即将 CarData 数据集中每个样本 7 个属性特征中的 3 个分配给参与方 A，并将标签特征 y 分配给 A，将其作为主动方；将剩下的 4 个特征分配给参与方 B，将其作为协作方。模型 FL-GRM 在 A，B 两方的共同参与下进行纵向联邦学习，实验结果如表 4 所列。

表 4　CarData 数据集上 2 个参与方的测试结果

Table 4　Results of 2 participants in CarData

模型	Deviance	Log-loss	AIC
LocalA-GRM	3162.50	9.94	18566.47
LocalB-GRM	3083.55	9.49	17580.73
NoFL-GRM	2569.87	9.23	17113.64
FL-LR	5957.44	9.52	17787.44
FL-PRM	45048.28	14.62	102455.98
FL-GRM	**2897.97**	**9.39**	**17403.86**

根据表 4 可以看出，CarData 数据集上 FL-GRM 的模型效果好于 LocalA-GRM 和 LocalB-GRM，说明联邦学习得到的模型比单方训练的模型更好。同时发现 FL-LR 的模型效果比 LocalA-GRM 更好，但 FL-GRM 的模型效果仍优于另外两种联邦学习模型，说明 FL-GRM 的拟合能力更强。

同时，本文在 CarData 数据集上使用有序洛伦兹曲线和 Gini 值来评价以上各个模型的拟合能力。图中紫色曲线是标签真实值排序分箱后的均值分布曲线，其他不同颜色的曲线是各模型的预测值排序分箱后的均值分布曲线。曲线越靠近紫色曲线，说明对应模型的拟合效果越好，同时越大的 Gini 值也代表模型拟合效果越好。首先将 FL-GRM 与本地模型进行对比，实验结果如图 3 所示。数据真实值的有序洛伦兹曲线对应着紫色曲线，可以看出，相比 LocalA-GRM 与 LocalB-GRM 对应的橙色和绿色曲线，NoFL-GRM 对应的褐色曲线整体更趋近于紫色曲线，说明 NoFL-GRM 的拟合效果更好。同时，FL-GRM 所对应的红色曲线也较为靠近紫色曲线，证明相比本地单方训练，联邦学习模型的拟合能力更强。图 4 是 FL-GRM 与两种联邦学习算法的对比实验结果，可以看出相比 FL-PRM 和 FL-LR 对应的橙色和绿色曲线，FL-GRM 对应的红色曲线更加趋近于紫色曲线，证明当标签特征分布为 Gamma 分布时，FL-GRM 的拟合能力优于 FL-LR 和 FL-PRM。

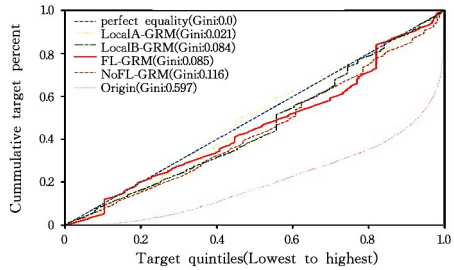

图 3　FL-GRM 与本地模型对比结果(电子版为彩图)

Fig. 3　Comparison results between FL-GRM and local models

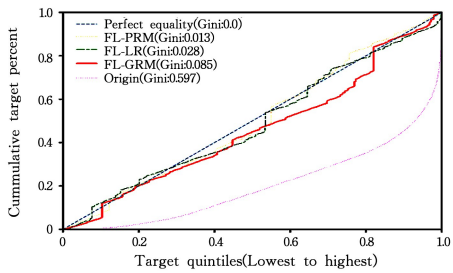

图 4　FL-GRM 与联邦模型对比结果

Fig. 4　Comparison results between FL-GRM and federated models

从图 3 和图 4 的实验结果可以发现，FL-GRM 得到的 Gini 值比 LocalA-GRM，LocalB-GRM，FL-PRM 和 FL-LR 都大，证明了 FL-GRM 的拟合能力更强。

图 5 是 FL-GRM 模型训练的损失值与迭代轮次之间的关系图，从图中可见在上述超参数条件下，本文提出的联邦 Gamma 回归模型参数更新方法能够稳定地沿着梯度降低的方向进行参数更新，使损失稳定下降，在迭代 240 轮左右模型可收敛。

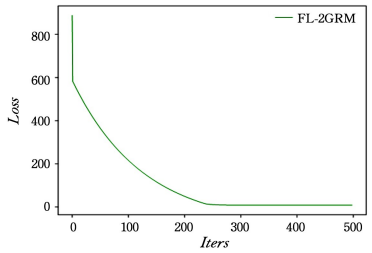

图 5 FL-GRM 迭代轮次与损失关系

Fig. 5 Relationship between rounds and losses

结束语 本文提出了一种基于联邦学习的 Gamma 回归算法，用于数据孤岛状态下多方联合进行 Gamma 回归模型训练。该算法应用迭代法推导出纵向联邦 Gamma 回归模型的对数似然估计式，并结合实际工程确定模型的连接函数，进而构造损失函数建立参数的梯度更新策略，最后对同态加密的各方参数进行融合更新，得到联邦 Gamma 回归模型。本文在两种数据集上进行的实验证明，应用联邦学习可以在隐私保护的前提下使用各方的数据集进行模型训练。同时模型测试结果证明，联邦学习的模型效果好于单方进行训练的效果；本文算法模型更新方式并不会因为参与方过多而产生精度损失；在标签特征服从 Gamma 分布的数据集中，联邦 Gamma 回归的模型效果强于联邦线性回归模型。

未来工作中我们将尝试建立其他分布函数对应的联邦回归模型，与标签满足不同分布函数的数据集相匹配，更好地利用各数据集的数据价值。

参 考 文 献

[1] NELDE R, JOHN A, ROBERT W W. Generalized linear models [J]. Journal of the Royal Statistical Society: Series A(General), 1972, 135(3): 370-384.

[2] ENGLAND P D, RICHARD J V. Stochastic claims reserving in general insurance[J]. British Actuarial Journal, 2002, 8(3): 443-518.

[3] AMIN M, QASIM M, AMANUM, et al. Performance of some ridge estimators for the gamma regression model[J]. Statistical Papers, 2020, 61(3): 997-1026.

[4] WU Z J. The Establishment of SPI_GD Drought index and the research on its test and application[D]. Lanzhou: Lanzhou University, 2017.

[5] MA X M, LUO Z Q. Human activity intensity time and space change research on haba snow mountain nature reserve[J]. Journal of Anhui Agricultural Sciences, 2015, 43(19): 205-208.

[6] PAYNTER S, NACHABE M. Regional scale spatio-temporal consistency of precipitation variables related to water resource management and planning[J]. Meteorological Applications, 2010, 16(3): 413-423.

[7] GONG M D. The empirical research on agricultural insurance-classification ratemaking[D]. Hunan: Hunan University, 2011.

[8] ZHONG Z, MENG S W. Comparison and application of gamma regression and lognormla regression[J]. Journal of Applied Statistics and Management, 2010, 29(3): 430-436.

[9] LI T, SANJABI M, BEIRAMI A, et al. Fair resource allocation in federated learning[J]. arXiv: 1905. 10497, 2022.

[10] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C]// Artificial Intelligence and Statistics. PMLR, 2017: 1273-1282.

[11] MCMAHAN H B, MOORE E, RAMAGE D, et al. Federated learning of deep networks using model averaging[J]. arXiv: 1602. 05629, 2016.

[12] GAO D, JU C, WEI X, et al. HHHFL: hierarchical heterogeneous horizontal federated learning for electroence phalography [J]. arXiv: 1909. 05784, 2019.

[13] LIU Y, KANG Y, ZHANG X, et al. A communication efficient vertical federated learning framework[J]. arXiv: 1912. 11187, 2019.

[14] SHREYA S, XING C, YANG L, et al. Secure and efficient federated transfer learning[C]. In 2019 IEEE International Conference on Big Data(Big Data). 2019: 2569-2576.

[15] HARD A, RAO K, MATHEWS R, et al. Federated learning for mobile keyboard prediction[J]. arXiv: 1811. 03604, 2018.

[16] LI T, ANIT K S, AMEET T, et al. Federated learning: Challenges, methods, and future directions[J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-60.

[17] KAIROUZ P, MCMAHAN H B, AVENT B. Advances and open problems in federated learning[J]. Foundations and Trends © in Machine Learning, 2021, 14(1/2): 1-210.

[18] MOTHUKURI V, PARIZI R M, POURIYEH S, et al. A survey on security and privacy of federated learning[J]. Future Generation Computer Systems, 2021, 115: 619-640.

[19] WANG J Z, KONG L W, HUANG Z C, et al. Summary of federated learning algorithms[J]. Big Data, 2020, 6(6): 64-82.

[20] YANG K, FAN T, CHEN T J, et al. A quasinewton method based vertical federated learning framework for logistic regression[J]. arXiv: 1912. 00513, 2019.

[21] YANG S, REN B, ZHOU X, et al. Parallel distributed logistic regression for vertical federated learning without thirdparty coordinator[J]. arXiv: 1911. 09824, 2019.

[22] LIU Y, LIU Y T, LIU Z, et al. Federated Forest[J]. arXiv: 1905. 10053, 2020.

[23] CHENG K, FAN T, JIN Y, et al. Secureboost: A lossless federated learning framework[J]. IEEE Intelligent Systems, 2021, 36(6): 87-98.

[24] YANG Q, LIU Y, CHEN T, et al. Federated machine learning: concept and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19.

[25] LI Q,WEN Z Y,WU Z M,et al. A survey on federated learning systems:vision,hype and reality for data privacy and protection [J]. arXiv:1907.09693,2022.

[26] JAKUB K H,BRENDAN M,DANIEL R,et al. Federated optimization:distributed machine learning for on-device intelligence [J]. arXiv:1610.02527,2022.

[27] WANG J Y,LIU Q H,LIANG H,et al. Tackling the objective inconsistency problem in heterogeneous federated optimization [C]// Advances in Neural Information Processing Systems. 2020:7611-7623.

[28] LUO X,WU Y,XIAO X,et al. Feature inference attack on model predictions in vertical federated learning[C]// International Conference on Data Engineering. 2021:181-192.

[29] WANC P,CHEN Q. Robust federated learning with attack-adaptive aggregation[J]. arXiv:2102.05257,2021.

[30] HAZRAT A,TANVIR A,MOWAFA H,et al. Federated Learning and Internet of Medical Things—Opportunities and Challenges[J]. Studies in Health Technology and Informatics,2022,295:201-204.

[31] WEN Y,CHEN M. Medical Data Sharing Scheme Combined with Federal Learning and Blockchain[J]. Computer Engineering,2022,48(5):145-153,161.

[32] WANG S S,CHEN J Y,LU Y N. COVID-19 chest CT image segmentation based on federated learning and blockchain[J]. Journal of Jilin University,2021,51(6):2164-2173.

[33] PINKAS B,SCHNEIDER T,ZOHNER M. Scalable private set intersection based on OT extension[J]. ACM Transactions on Privacy and Security(TOPS),2018,21(2):1-35.

[34] TANG Y Y. Gamma distribution and gamma regression[D]. Yangzhou:Yangzhou University,2017.

[35] RIVEST R L,ADLEMAN L M,DERTOUZOS M L. On data banks and privacy homomorphisms[J]. Foundations of secure computation,1978,4(11):169-180.

GUO Yan-qing, born in 1980, Ph.D, professor, Ph.D supervisor. His main research interests include machine learning, computer vision and cyberspace security.

FU Hai-yan, born in 1981, Ph.D, senior engineer. Her main research interests include federated learning, image retrieval and computer vision.

(责任编辑:何杨)